April 2022

On the Robustness of Argumentative Explanations
Supplementary Material

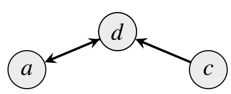
A. Cut-off dispute trees and defense sets (Proofs of Section 3)

Lemma 3.4. *It holds that (a) each cut-off dispute tree is finite; and (b) for each dispute tree* $\mathtt{t}$*, there is a unique cut-off dispute tree* $\mathtt{t}'$*, and* $D(\mathtt{t}') = D(\mathtt{t})$ *(but not vice versa).*

Proof. (1) holds by definition of the cut-off dispute tree (and by assumption A is finite).

(2) By definition, each root-to-leaf path in a dispute tree is cut off after the first repetition; therefore, the corresponding cut-off tree is uniquely defined. For the other direction, we can construct a counter-example:

F :



$$\mathtt{t}_1 : \quad [P:a] \begin{cases} [O:d] — [P:c] \\ [O:b] — [P:a] — [O:b] — [P:c] \end{cases}$$

$$\mathtt{t}_2 : \quad [P:a] \begin{cases} [O:d] — [P:c] \\ [O:b] — [P:a] — [O:b] \cdots \end{cases}$$

The trees $\mathtt{t}_1$ and $\mathtt{t}_2$ yield the same cut-off tree. Furthermore note that both trees correspond to a minimal defense set (therefore, minimality of the defense set does not guarantee uniqueness of the corresponding dispute tree either). $\square$

Proposition 3.5. *Let* $F = (A,R)$ *be an AF and let* $\mathtt{t}$ *be an admissible cut-off tree. Then* $\mathsf{D}(\mathtt{t})$ *is admissible in* F*; and there exists an admissible dispute tree* $\mathtt{t}'$ *with* $D(\mathtt{t}) = D(\mathtt{t}')$*.*

Proof. First, we show that $D(\mathtt{t})$ is admissible. Let $S = \{x \mid \exists [P:x]_i \in \mathtt{t}\}$ denote the set of all proponent arguments. Note that the tree $\mathtt{t}$ contains all attacker of S in F. The set S is conflict-free; otherwise, $\mathtt{t}$ would contain an argument labelled with P and O, contradiction. Moreover, S defends itself because each opponent is attacked.

It remains to show that we can construct an admissible dispute tree $\mathtt{t}'$ with $D(\mathtt{t}) = \mathsf{D}(\mathtt{t}')$. We can extend each cut-off tree by extending each root-to-leaf path in a loop whenever an argument is repeated. That is, we extend each root-to-leaf path B containing a repetition $[P:a]_i \dots [P:a]_k$, $i < k$, infinitely many times with the sub-tree with root $[P:a]_i$. The so obtained tree $\mathtt{t}'$ does not contain arguments labelled with both P and O since we only copied nodes that already existed in $\mathtt{t}$; moreover, no new arguments have been introduced. We obtain that $\mathtt{t}'$ is admissible and $D(\mathtt{t}) = \mathsf{D}(\mathtt{t}')$. $\square$

Let us consider an example that that shows that for cut-off trees, an exponential blow-up cannot be avoided.

Example A.1. *An AF with an exponential cut-off tree can be constructed as follows: we consider n proponent arguments* $a_1, \dots, a_n$*; each* a_i *is attacked by n attackers* $x^i_1, \dots, x^i_n$*; moreover, each argument* x^i_j *is attacked by argument* a_k *where* $k = (i + j)$ *mod n (the argument* a_n *attacks each* x^i_j *with* $(i + j)$ *mod* $n = 0$*). In total, the AF has* $n^2 + n$ *many arguments but the dispute trees for* a_i *are exponential: each argument* a_i *is attacked by n arguments; and in each kth opponent (even) level of the tree, we encounter at most* $n - k$ *repetitions in the paths (therefore, we cut off at most* $n - k$ *paths in the kth level). The dispute tree* $n(n-1)\dots(n-k)$ *many opponent arguments in the kth level and is therefore exponential in the number of arguments.*

Proposition 3.7. *The explanation method $expl_{\mathtt{t}}$ satisfies **(F)** but not **(C_σ)**, **(T)** and **(M)**.*

Proof. Finiteness is satisfied by definition. Example A.1 is a counter-example for comprehensibility; it is not tractable unless P=NP(if a dispute tree can be computed in Pthen the credulous acceptance problem for admissible semantics can be solved in P); and it is not minimal. □

Proposition 3.9. *The explanation method $expl_D$ satisfies **(F)** and **(C_σ)** for $\sigma \in \{ad, pr, co\}$; it does not satisfy **(T)** and **(M)**.*

Proof. Finiteness is satisfied iff the underlying AF is finite; comprehensibility is satisfied because checking admissibility is in P. □

Proposition 3.11. *The explanation method $expl_D^{\min}$ satisfies **(F)**, **(M)**, and **(C_σ)** for $\sigma \in \{ad, co, pr\}$; it does not satisfy **(T)**.*

Proof. Minimality is satisfied by definition; finiteness is satisfied iff the underlying AF is finite; comprehensibility is satisfied because checking admissibility is in P. □

B. Robustness and Reliability (Proofs of Section 4)

Proposition 4.2. *The explanation method $expl_D^{\min}$ satisfies **(R_{ad})**.*

Proof. Consider two AFs F and F' with $ad(F) = ad(AF')$ and consider a common argument a. Let $\mathtt{t}$ be a minimal dispute tree for a in F. We show that we can construct a minimal dispute tree $\mathtt{t}'$ for a in F' using $D(\mathtt{t})$.

The set $E = D(\mathtt{t})$ is admissible in F; hence, $D(\mathtt{t})$ is admissible in F' by assumption. By Proposition 2.5 (b), we can construct a dispute tree $\mathtt{t}'$ for a in F' from arguments in E. It remains to show that $D(\mathtt{t}') = E$ and $\mathtt{t}'$ is a minimal dispute tree for a in F': Towards a contradiction, suppose that $D(\mathtt{t}') \subsetneq E$. Then $E' = D(\mathtt{t}')$ is admissible in F'; hence, by assumption, E' is admissible in F. Moreover, $a \in E'$. By Proposition 2.5 (b), we can construct a dispute tree $\mathtt{t}''$ from arguments in E'. But then $D(\mathtt{t})$ is a proper superset of $D(\mathtt{t}'')$. We obtain a contradiction to the minimality of $\mathtt{t}$ in F. □

Proposition 4.4. *The explanation methods $expl_{\mathtt{t}}$ and $expl_D$ do not satisfy **(R_σ)** for any semantics σ under consideration; $expl_D^{\min}$ does not satisfy **(R_σ)** for $\sigma \in \{co, pr\}$.*

Proof. See Example 4.3 and 4.1. □

Proposition 4.5. *The explanation methods $expl_{\mathtt{t}}$, $expl_D$ and $expl_D^{\min}$ satisfy **(Rel_σ)** for $\sigma \in \{ad, co, pr\}$.*

Proof. We obtain σ-reliabiltiy since credulous acceptance for admissible semantics is preserved for the considered semantics. Let $\mathtt{t}$ be a dispute tree for a in F. The set $D(\mathtt{t})$ is admissible, i.e., $D(\mathtt{t}) \in ad(F)$ and there is $S \in \sigma(F)$ with $D(\mathtt{t}) \subseteq S$, for $\sigma \in \{pr, co\}$. We can construct a dispute tree for a from S in F'. □

Proposition 4.9. *The explanation method $expl_D^{\min}$ satisfies **(SR_σ)** for $\sigma \in \{co, ad, pr\}$; moreover, $expl_{\mathtt{t}}$, $expl_D$ and $expl_{\mathtt{t}}^{\min}$ satisfy **(SR_{co})**. The explanation methods $expl_{\mathtt{t}}$, $expl_D$ and $expl_{\mathtt{t}}^{\min}$ do not satisfy **(SR_σ)** for $\sigma \in \{ad, pr\}$.*

Proof. First, we show that all explanation methods satisfy **(SR$_{co}$)**.

Let $F = (A,R)$, $F' = (A',R')$ be two strongly equivalent AFs wrt semantics σ, and let $a \in A \cup A'$.

- We show that $expl_{\mathtt{t}}$ is co-robust. Consider an explanation (dispute tree) $\mathtt{t} = expl_{\mathtt{t}}(F,a)$. Since we delete only attacks between self-attacker in the complete kernel F^{ck}, we have that $\mathtt{t}$ is an explanation for a in F^{ck}. Since by assumption $F^{ck} = (F')^{ck}$, we obtain that $\mathtt{t}$ is an explanation for a in F'.
- Since the dispute trees in both AFs correspond to each other, it follows that $expl_D$ and $expl_D^{\min}$ are co-robust.

Next, we show that $expl_D^{\min}$ satisfies strong σ-robustness for $\sigma \in \{ad, pr\}$. Let $F = (A,R)$, $F' = (A',R')$ be two strongly equivalent AFs wrt semantics σ, and let $a \in A \cup A'$. It holds that $ad(F) = ad(F')$ (since their admissible kernel is syntactically equivalent). Therefore, we obtain $expl_D^{\min}(F,a) = expl_D^{\min}(F',a)$ from Proposition 4.2.

We can construct counter-examples for the remaining cases. □

Proposition 4.10. *The explanation methods $expl_D$, $expl_{\mathtt{t}}$ and $expl_D^{\min}$ satisfy **(SRel$_\sigma$)** for $\sigma \in \{ad, co, pr\}$.*

Proof. Analogous to the proof for Proposition 4.5, we can infer σ-reliabiltiy for admissible, preferred, and complete semantics since credulous acceptance for admissible semantics is preserved. □

Strong gr-realiablity We provide an example which shows that no explanation method is reliable under gr semantics.

Example B.1. *Consider an AF $F = (\{a,b\}, \{(a,b),(b,a),(b,b)\})$. Recall that the grounded kernel $F^{gk} = (A, R^{gk})$ of an AF $F = (A,R)$ is defined with $R^{gk} = R \setminus \{(a,b) \mid a \neq b, (b,b) \in R, \{(b,a),(a,a)\} \cap R \neq \emptyset\}$. In F, a has an admissible dispute tree whereas in F^{gk}, the attack (a,b) is deleted which renders a unacceptable in the grounded kernel.*

C. Pseudo-robustness (Proofs of Section 5)

Lemma C.1. *Given $F = (A,R)$, an argument $a \in A$ and a set $S \in ad(F)$ with $a \in S$. Constructing a defense set is in* P.

Proposition 5.1. *The explanation methods $expl_D$ and $expl_D^{\min}$ satisfy **(PR$_\sigma$)** for $\sigma \in \{ad, co\}$.*

Proof. Analogous to the proof of [18, Proposition 4.1] we can construct a $\subseteq$-minimal set that contains a and is contained in S. Since the $\subseteq$-minimal admissible sets containing a correspond to $Expl_D^{\min}(F,a)$ the statement follows. □

Proof. Consider two AFs F and F' with $\sigma(F) = \sigma(F')$, and let S denote a defense set for some argument a in F.

- $\sigma = ad$: By Proposition 3.5, S is admissible in F'. By Lemma C.1, we can construct a defense set for a in F' in polynomial time.

- $\sigma = co$: Now, we cannot assume that the given defense set is admissible in F'. Given S, we construct the complete set S' that contains S by iteratively adding all arguments that are defended by S in F until a fixed point is reached. The set S' is complete in F and therefore also in F'. We proceed as in the case for admissible semantics to obtain the defense set for a in F'.

□

Proposition 5.3. *The explanation methods $expl_D$ and $expl_D^{\min}$ satisfy* **(SPS$_\sigma$)** *for* $\sigma \in \{ad, co, pr\}$.

Proof. Complete semantics guarantees update tractability under strong equivalence for our considered explanation types since they are *co*-robust; strong equivalence under preferred semantics guarantees equivalence of admissible semantics (recall that two AFs are strongly equivalent to each other w.r.t. preferred semantics iff the admissible kernel coincides). □